Hospital Early Readmission Classifier

Alexander Gaujean & Rima Bazerji

Problem

Hospital readmission after an admitted patient is discharged is a high priority for hospitals. In total Hospital Readmissions are one of the most costly episodes to treat, costing Medicare about **$26 billion annually**, with about **$17 billion consisting of avoidable trips after discharge**.

The health care burden of hospitalized patients with diabetes (1 and 2) is substantial and only growing. As of today, around 9% of Americans have diabetes or prediabetes, according to a recent CDC report.

A better understanding of the factors that lead to hospital readmissions could help decision makers understand potential ways to reduce early readmissions (within 30 days) and provide more efficient care.

Classification problem : We focused our project on attempting to predict whether a discharged diabetic patient will be readmitted into a hospital within 30 days (Target = 1) or not (Target = 0).

Data (‘diabetic_data.csv’)

	encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	...	citog
0	2278392	8222157	Caucasian	Female	[0-10)	?	6	25	1	1	...	
1	149190	55629189	Caucasian	Female	[10-20)	?	1	1	7	3	...	
2	64410	86047875	AfricanAmerican	Female	[20-30)	?	1	1	7	2	...	
3	500364	82442376	Caucasian	Male	[30-40)	?	1	1	7	2	...	
4	16680	42519267	Caucasian	Male	[40-50)	?	1	1	7	1	...	

Diabetic Patient Data from 130 U.S Hospitals (1998-2008)

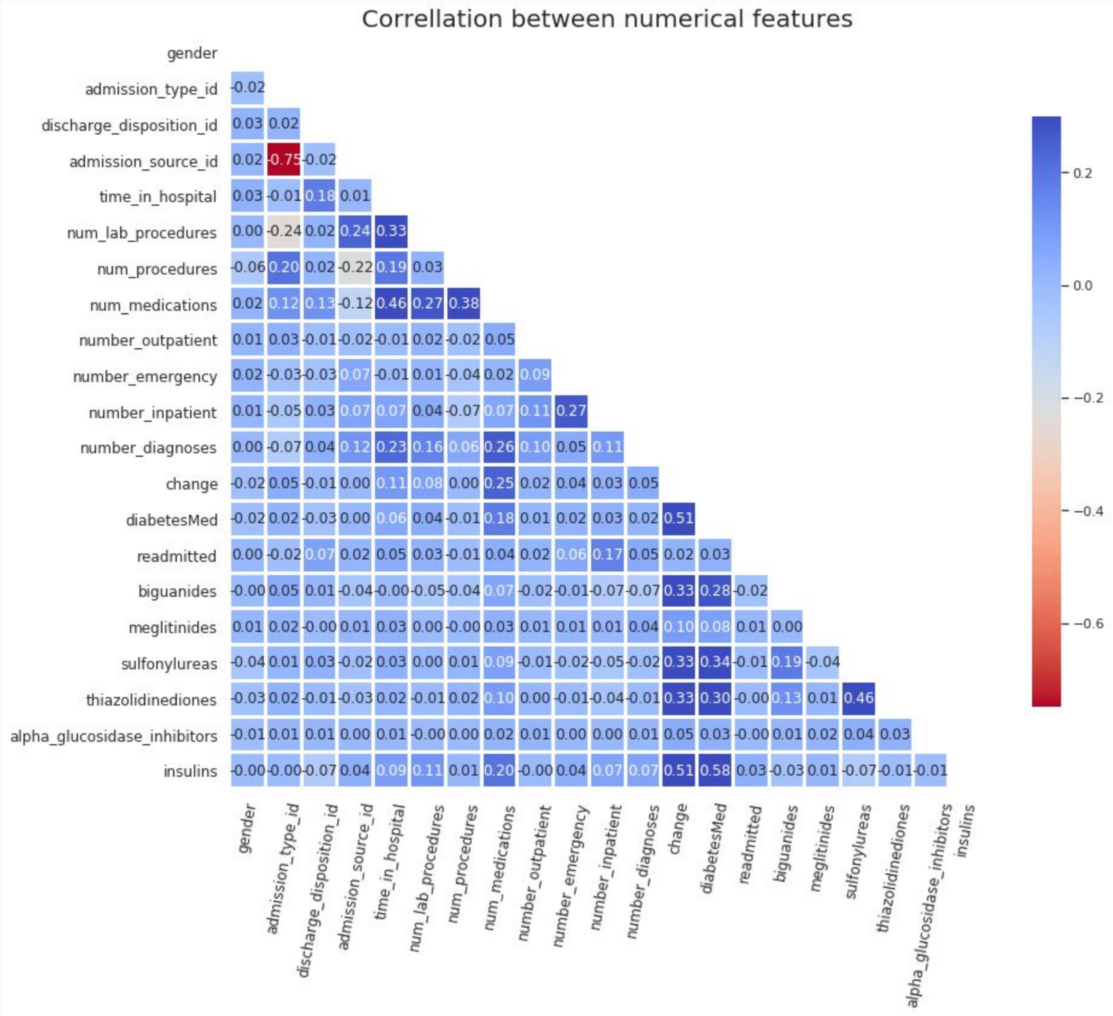
citoglipton	insulin	glyburide-metformin	glipizide-metformin	glimepiride-pioglitazone	metformin-rosiglitazone	metformin-pioglitazone	change	diabetesMed	readmitted
No	No	No	No	No	No	No	No	No	NO
No	Up	No	No	No	No	No	Ch	Yes	>30
No	No	No	No	No	No	No	No	Yes	NO
No	Up	No	No	No	No	No	Ch	Yes	NO
No	Steady	No	No	No	No	No	Ch	Yes	NO

Sources:

- UCI Machine Learning.com Diabetes dataset
- Webscraping ICD9 codes from http://www.icd9data.com/

Correlation between numerical Features

- Slight correlation between **admission source id** (physician referral, emergency room, and transfer from a hospital) & **admission type id** (emergency/newborn etc)



Feature Engineering

Categorical Data

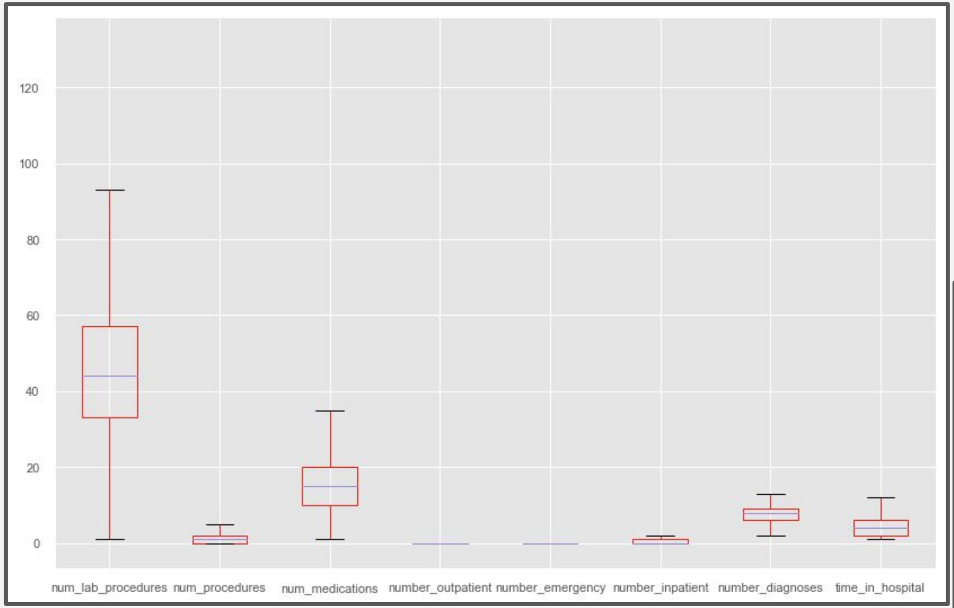
- Created bins:
  - Medication grouped by drug class
  - ICD-9 Codes
  - Grouped 3 diagnosis into boolean predictor
  - Retained top 10 Medical Specialties
- Dummy variables

Web scraped and formatted the ICD-9 codes in order to group our columns

	icd9_range	diagnosis_desc
0	[001, 139]	Infectious And Parasitic Diseases
1	[140, 239]	Neoplasms
2	[240, 279]	Endocrine, Nutritional And Metabolic Diseases,...
3	[280, 289]	Diseases Of The Blood And Blood-Forming Organs
4	[290, 319]	Mental Disorders
5	[320, 389]	Diseases Of The Nervous System And Sense Organs
6	[390, 459]	Diseases Of The Circulatory System
7	[460, 519]	Diseases Of The Respiratory System
8	[520, 579]	Diseases Of The Digestive System
9	[580, 629]	Diseases Of The Genitourinary System
10	[630, 677]	Complications Of Pregnancy, Childbirth, And Th...

```
df['alpha_glucosidase_inhibitors'] = df[(df['acarbose'] == True) | (df['miglitol'] == True)].any(axis=1)
df['alpha_glucosidase_inhibitors'] = df['alpha_glucosidase_inhibitors'].fillna(False)
```

PCA for Continuous Variables - Dimensionality Reduction



Continuous Variables Before Scaling and PCA Transformation

Continuous Variables Reduced to Six Principal Components (cutoff was an explained_variance_ratio of .075)

Class Imbalance & Train/Test split

- Class Imbalance:

  Readmitted patients accounted for ~ 11% of total

  ⇨ Undersampled the non readmitted patients to achieve 1:1 ratio

  ⇨ Obtained a set of 10,625 values for both our Target Values

- Train-Test-Split :

  80% train~ 8,500 patients

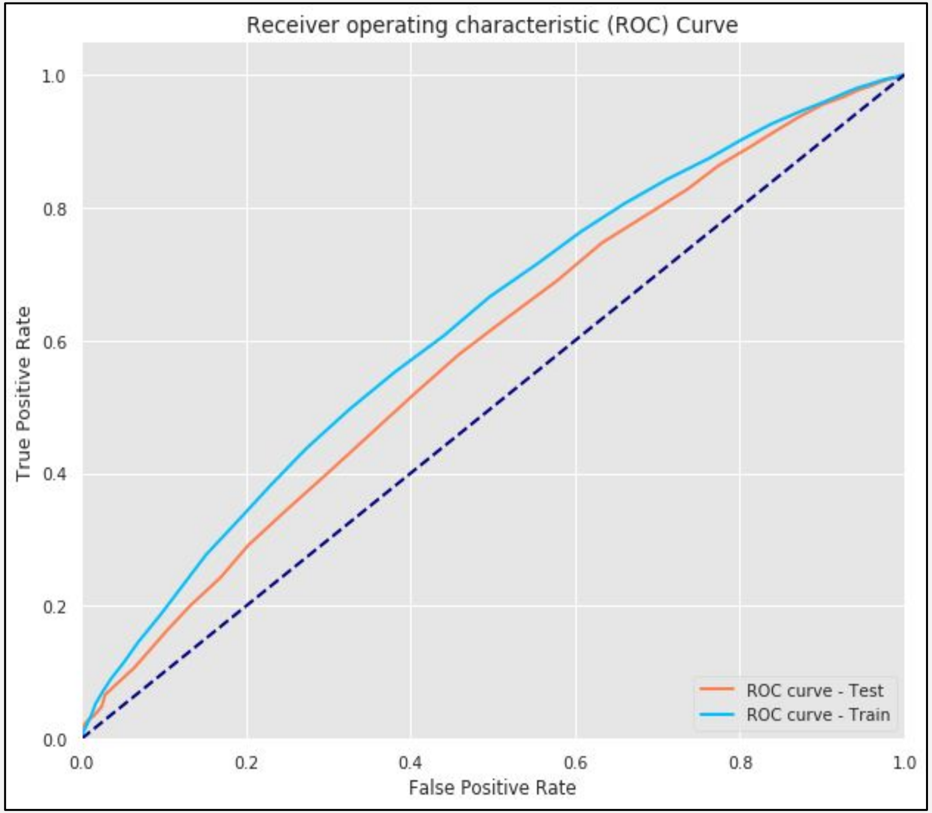
Baseline Model: Logistic regression

- No tuning of parameters

  - accuracy train score: 0.6142839746712129
  - accuracy test score: 0.6042377009254749
  - recall score for test: **0.5366327025715673**
  - recall score for train: **0.5548512920526573**
  - precision score for test: 0.6227477477477478
  - precision score for train: 0.6291637871458189

Model: KNN

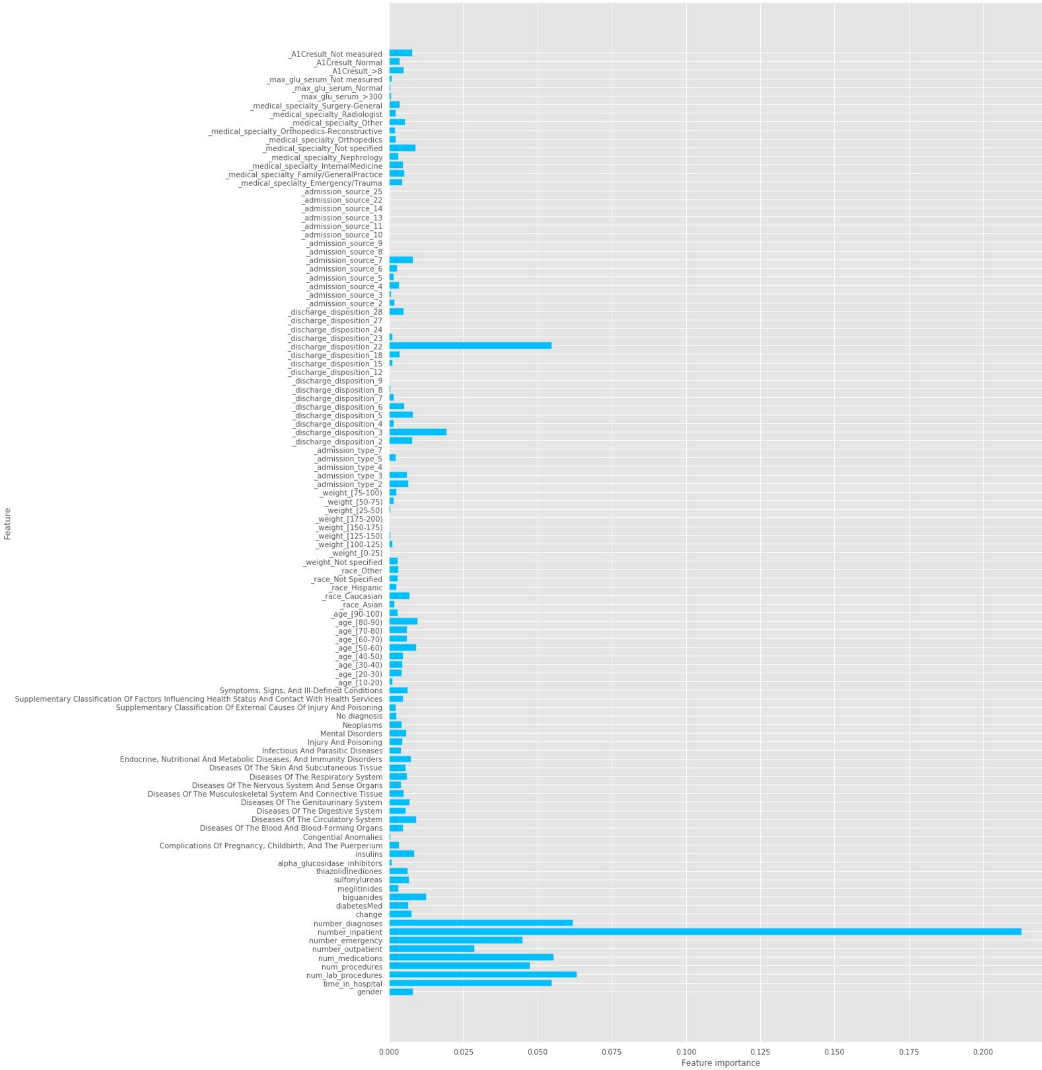
- Computationally expensive to identify the best value of K
- Doesn't learn from training
- Poor option for our data because it is susceptible to noise



Model: Random Forest

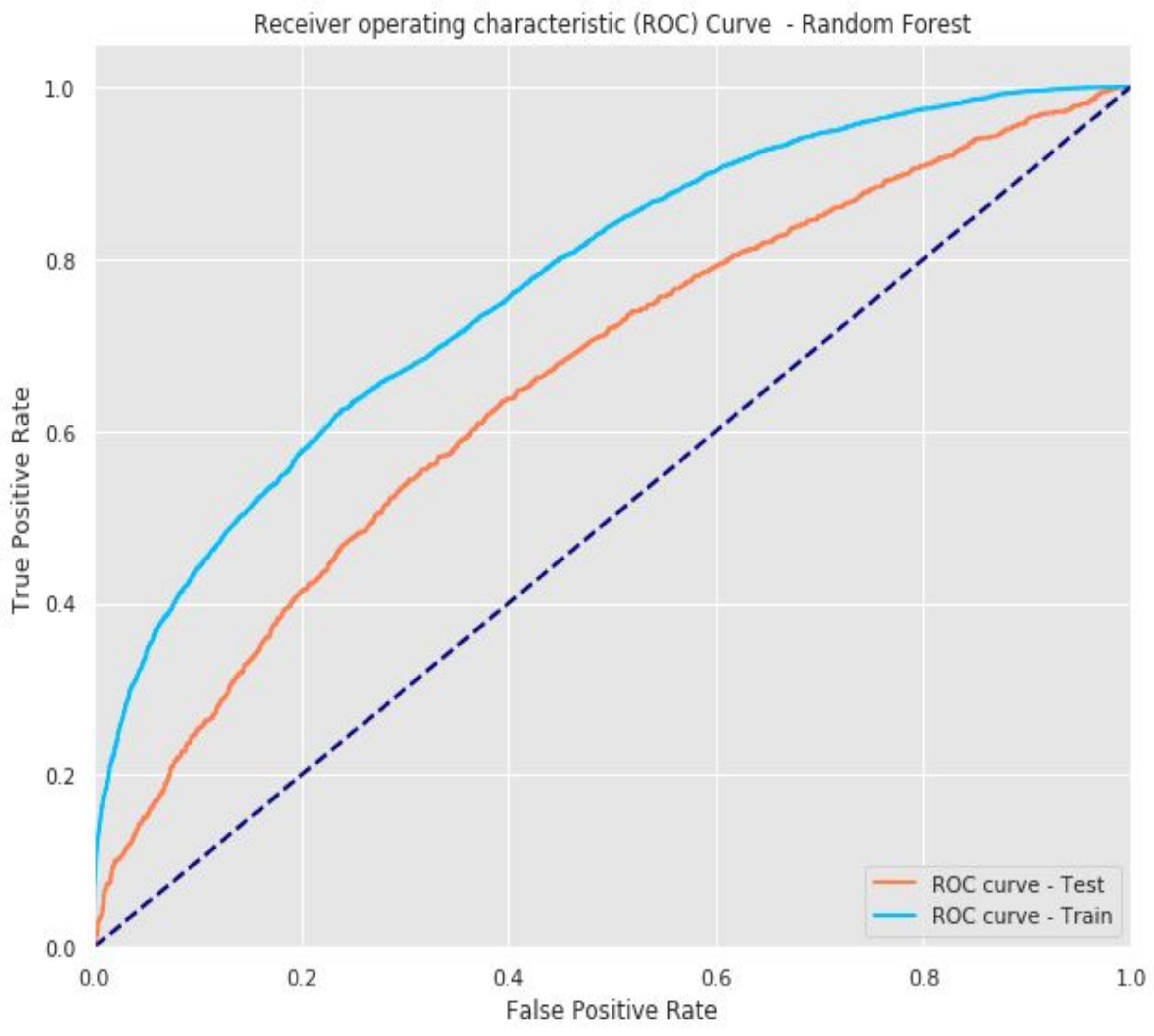
Most Important Features:

- Number of Inpatient visits within the last year
- Discharge disp #22 : “Transfer to rehab facility”
- Number of emergency visits
- Number of lab procedures
- Time in hospital (in days)
- Number of Procedures
- Discharge disp #3: “Transfer to Skilled nursing facility”



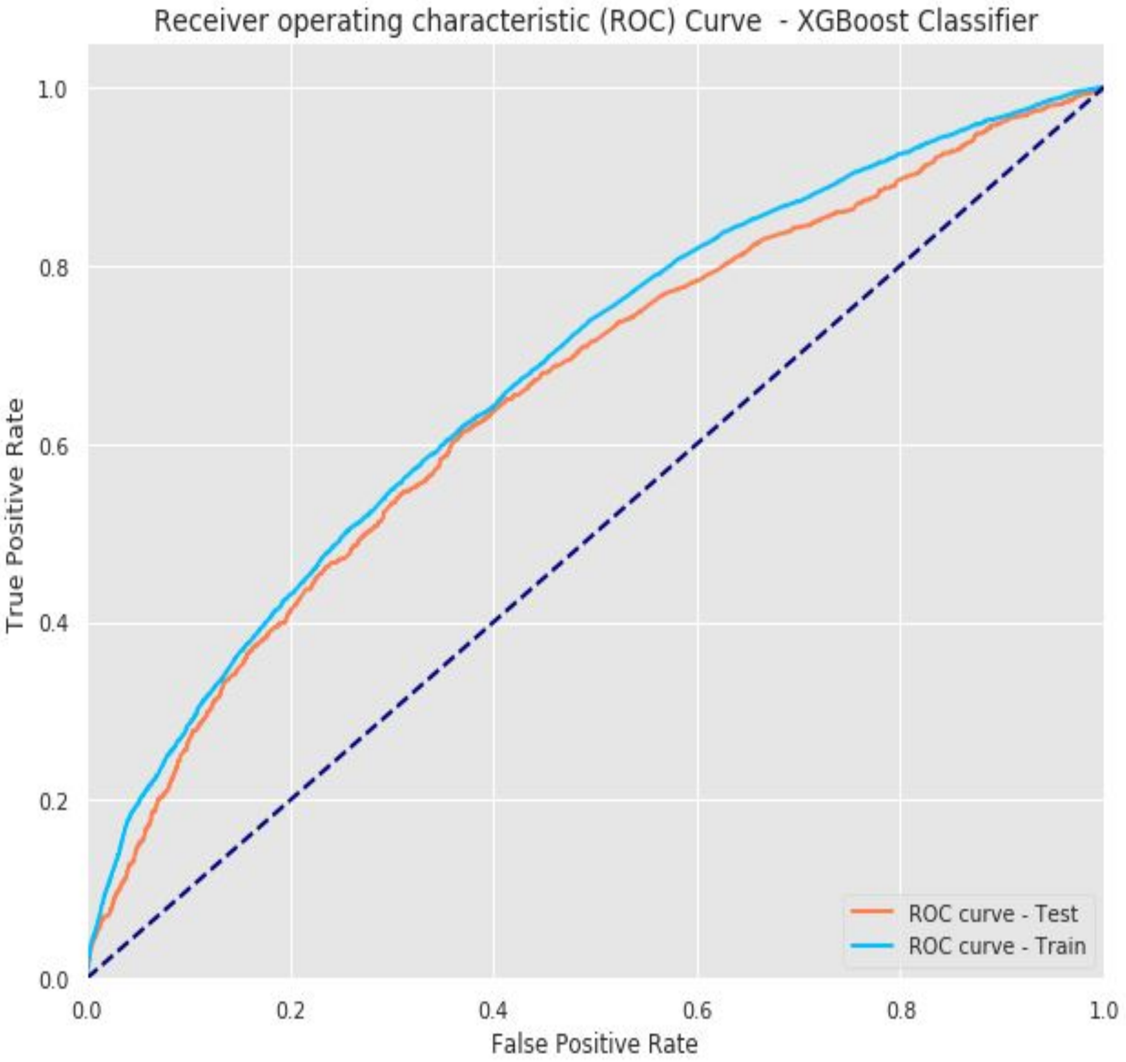
Model: Random Forest

- Significant difference between our test and train data
- Performed well on our training data



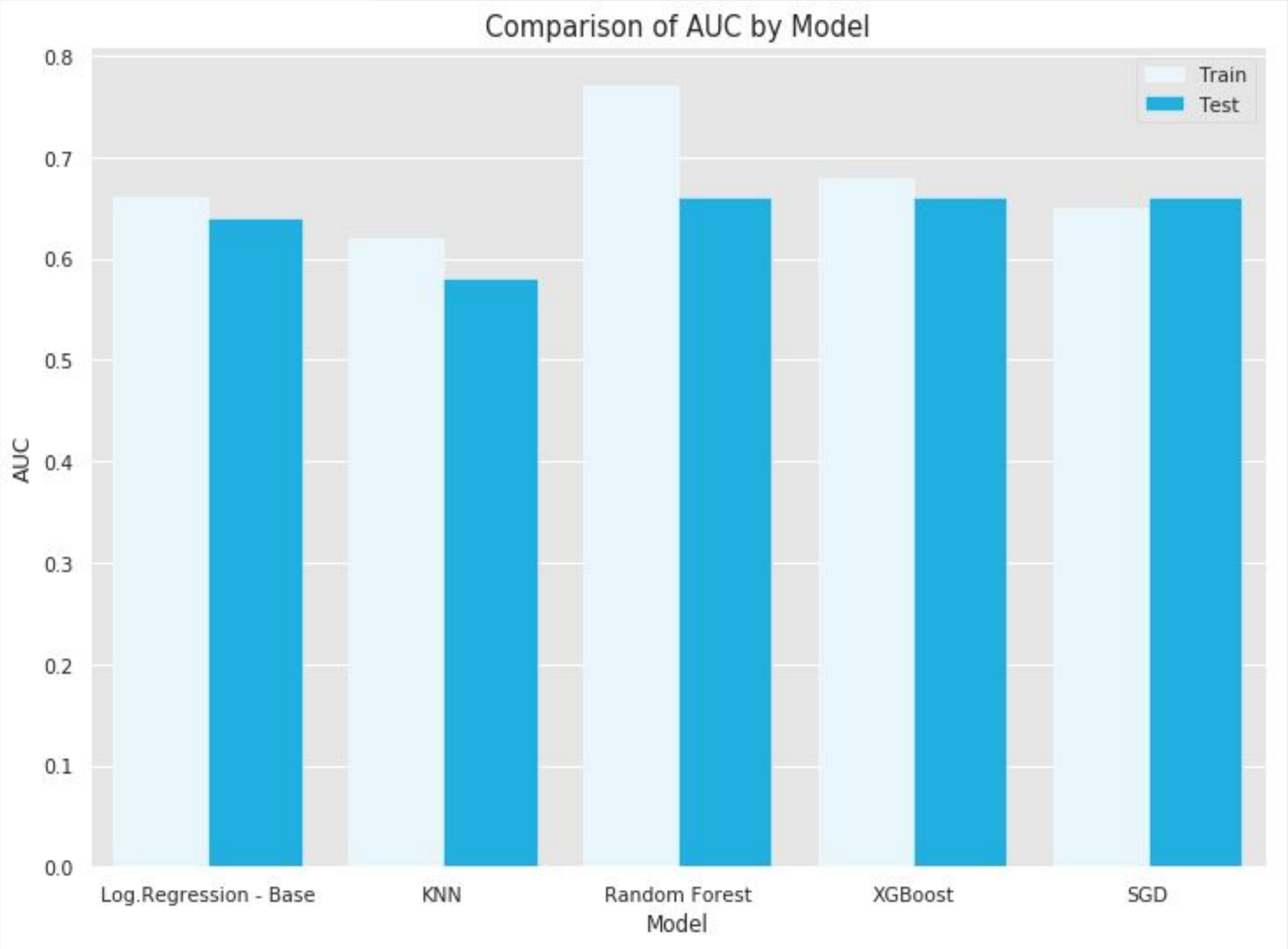
Model: XGBoost

- Able to slightly improve the way our model generalized to new information.
- Not perfectly predictable but XGBoost performed better when we reduced our dimensionality.



ROC - AUC

- Our Random Forest model had the highest train score.
- XGBoost has an overall better area under the curve for our testing data.



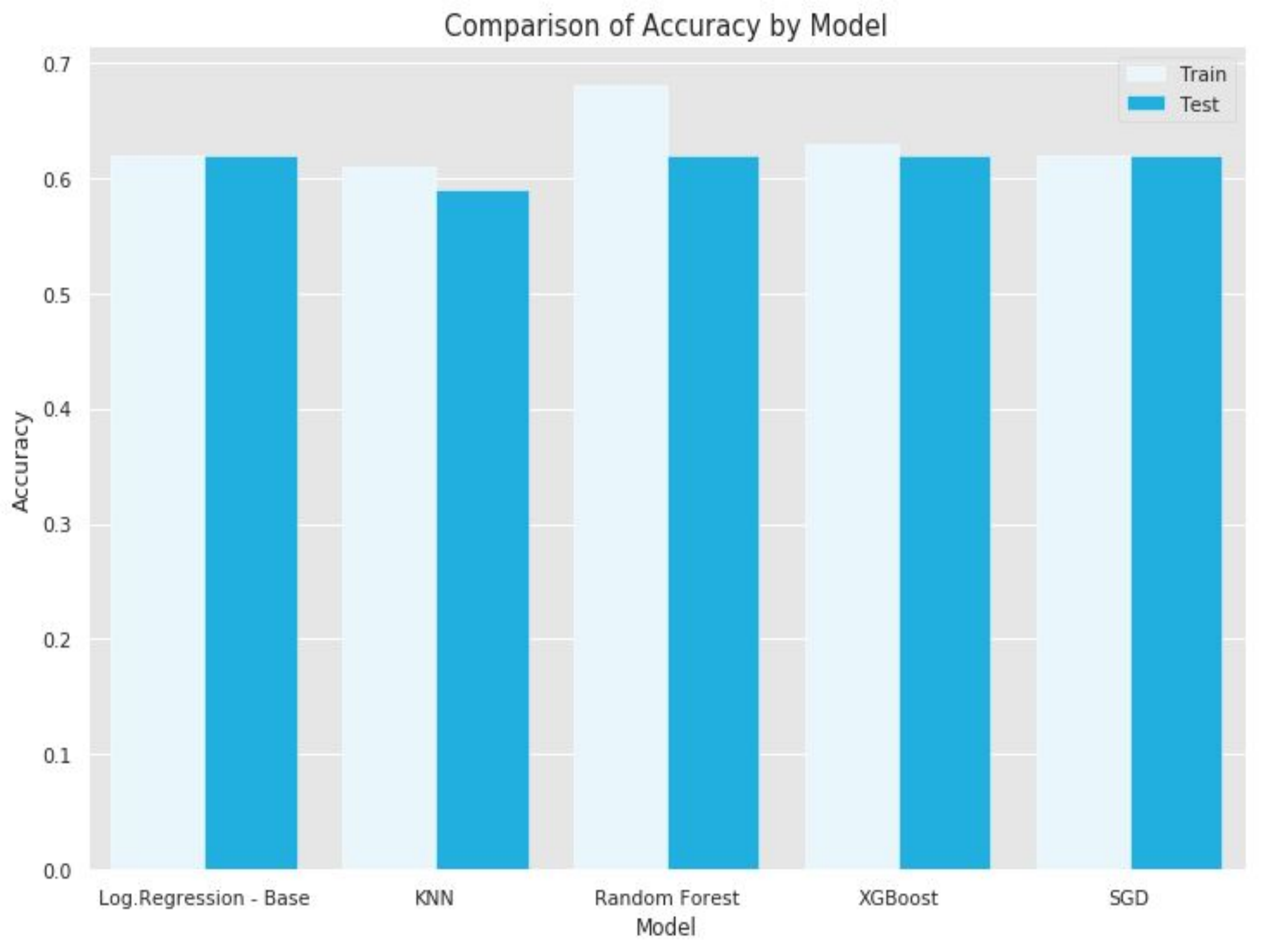
Accuracy

Top Performing Models:

Random Forest:
68.4 % Train
61.7 % Test

VS

XG Boost:
62.5 % Train
62.1 % Test



Hyperparameter tuning using GridSearch

MODEL	Hyperparameters Grid	Best parameters	Recall
Logistic Regression	{"max_iter" : [100,200,500], "penalty": ['L1', 'L2'] , "C" : [1,10]}	{"max_iter" : [100], "penalty": ["L2'] , "C" : [10]} (Ridge)	Train: 55 % Test: 53 %
KNN	{'n_neighbors' : [3,5, 100,300]}	{'n_neighbors' : [100]}	Train: 50 % Test: 48 %
Random Forest	{'criterion' : ['gini','entropy'], 'max_depth': [3,5,10,20], 'n_estimators': [100, 115, 150]}	{‘criterion’ = 'entropy', ‘max_depth’ = 10, ‘n_estimators’ = 100}	Train: 64% Test: 58%
XGBoost	{'max_depth': [3,10], 'learning_rate' : [0.001, 0.16, 0.2], 'gamma' : [3,7,10], 'n_estimators': [100]}	{'gamma': 7, 'learning_rate': 0.16, 'max_depth': 3, 'n_estimators': 100}	Train: 62% Test: 61%

Our Discoveries

- For our Classifier we considered the most important evaluation metric to be **recall.**
- In order to help hospital administration effectively address the issue of Hospital Readmissions <30 days we want our model to be able to correctly predict as many of the relevant cases as possible.
- No real surprise that the *number of previous inpatient visits* proved to be the strongest predictor of readmission within 30 days.
- Another strong predictor were discharge dispositions #22 and # 3, which correspond to discharge to a rehab facility or skilled nursing facility, respectively.
- Patients who were on medications that fell into the ‘biguanides’ group showed a propensity to readmission.

Conclusion

Model improvements:

- Feature Engineering
- Other models such as Support Vector Machine
- Multiclass with three possible classes (<30, >30, No)

Recommendations:

- The most important indicator of readmission is the *number of previous inpatient admissions*, which could speak to the idea that an individual's lifestyle choices and not simply their ailments must be addressed before and after release.
- Releasing a patient for continued treatment and rehabilitation might seem proactive and safe, however our data shows that discharge to these facilities prove to have no positive impact on overall health improvement.
- A final thought was that the number of lab procedures proved to be more predictive than the number of diagnoses when identifying a patient who was going to be readmitted. This could be because lab procedures indicate a complicated diagnoses, but it also highlights the inefficient and expensive approach healthcare facilities take when treating patients.